MiPV AI-IR at PAN 2026: Not Just Black Boxes, Interpretable AI-Generated Text Detection with Heterogeneous Signals*

Notebook for the PAN Lab at CLEF 2026

Teresa Vaccari[1,2], Pelinsu Topaloglu[1,2], Artem Avdeiko[1,2] and Georgios Peikos[2,*]

[1]University of Pavia, 27100 Pavia, Italy
[2]University of Milano-Bicocca (DISCo), 20126 Milan, Italy

Abstract

This paper describes the systems submitted by the MiPV AI-IR team to the third Voight-Kampff Generative AI/LLM Detection Task at PAN and ELOQUENT 2026. The task focuses on binary AI-generated text detection, where each text must be classified as either human-authored or machine-authored under challenging style-mimicking conditions. We propose a heterogeneous and partially interpretable ensemble that combines lexical, linguistic, probabilistic, and semantic signals. The system includes a TF-IDF lexical classifier, a handcrafted linguistic feature classifier, a probabilistic detector based on Binoculars and related language-modeling features, and a fine-tuned DistilBERT classifier. Their probability scores are combined through a logistic regression aggregation layer. Validation results show that lexical and semantic subsystems achieve the strongest standalone performance, while linguistic features provide interpretable genre-sensitive signals. The aggregation layer obtains the best validation performance, suggesting that combining heterogeneous evidence can provide a modest improvement over individual subsystems.

Keywords

AI-generated text detection, LLM detection, Machine-authored text, Interpretable AI

1. Introduction

The use of Artificial Intelligence (AI) for text generation has increased rapidly, with Large Language Models (LLMs) now being adopted in several application contexts, including education, professional writing, journalism, and online communication [1, 2, 3]. At the same time, the quality and fluency of AI-generated text have improved substantially, making it increasingly difficult to distinguish machine-authored content from human-written content [4]. This growing realism creates important challenges for authorship verification [5], misinformation detection [6], content authenticity [7], and the broader assessment of information trustworthiness. As a result, developing reliable methods for identifying AI-generated text has become increasingly important across academic, professional, and societal contexts.

AI-generated text detection becomes substantially more challenging when machine-authored texts are produced under instructions that explicitly ask LLMs to imitate the writing style of specific human authors or to reduce detectable machine-authored patterns. This challenge is likely to become more pronounced as users become increasingly familiar with prompt engineering strategies that make AI-generated text more human-like and harder to identify. As a result, robust AI-generated text detection requires systems capable of **leveraging multiple complementary signals** that remain informative under challenging and evolving generation conditions.

Moreover, AI-generated text detection systems are increasingly being used in contexts where their decisions may directly affect individuals. For example, detection systems may be used to determine whether

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
teresa.vaccari01@universitadipavia.it (T. Vaccari); pelinsu.topaloglu01@universitadipavia.it (P. Topaloglu); artem.avdeiko01@universitadipavia.it (A. Avdeiko); georgios.peikos@unimib.it (G. Peikos)
0009-0000-6696-2374 (T. Vaccari); 0009-0002-3690-2463 (P. Topaloglu); 0009-0004-7448-9272 (A. Avdeiko); 0000-0002-2862-8209 (G. Peikos)

a student assignment, professional report, or published article was partially or entirely generated using AI tools. Recent public controversies have also highlighted cases in which AI-authorship accusations or automated detection outcomes affected professional collaborations and freelance work [8, 9]. In such settings, detection outcomes may lead to disputes, reputational consequences, or professional sanctions. As a result, relying exclusively on black-box systems becomes problematic, since authors, institutions, and readers may require clear explanations regarding why a text was classified as machine-authored. **Interpretable detection approaches** can therefore support transparency, human review, and more informed decision-making processes by providing insight into the signals and patterns that contributed to the final prediction.

Together, these challenges motivate the development of AI-generated text detection systems that are both robust and partially interpretable. In this work, we investigate whether heterogeneous detection systems can support robust AI-generated text detection under adversarial and style-mimicking conditions. Rather than relying exclusively on a single system, we combine lexical, linguistic, probabilistic, and semantic subsystems designed to capture complementary properties of machine-authored text.

Recent research on AI-generated text detection has explored a broad range of approaches, including lexical and stylometric methods [10, 11, 12], probabilistic zero-shot detectors [13, 14, 15], transformer-based classifiers [16, 17, 18, 19], and some hybrid ensemble systems [20, 21, 22]. While transformer-based architectures often achieve strong predictive performance, lightweight lexical and probabilistic approaches offer advantages due to their efficiency, interpretability, and reduced training requirements.

For this reason, our approach investigates a heterogeneous ensemble architecture that combines multiple types of evidence for AI-generated text detection. The goal of this paper is to examine whether a partially interpretable ensemble can support robust detection across different genres and generation conditions while also providing insight into the signals that contribute to the final prediction. The proposed system combines four complementary subsystems:

- a lexical subsystem based on TF-IDF representations over word and character n-grams;
- a linguistic subsystem based on handcrafted stylometric and discourse-oriented features inspired by linguistic analyses of LLM-generated text;
- a probabilistic subsystem based on the Binoculars score and additional language-modeling features, including surprisal, entropy, and log-rank;
- a semantic subsystem based on a fine-tuned DistilBERT model for capturing contextual and semantic patterns.

The probability scores produced by these subsystems are combined through a logistic regression aggregation layer. This design allows us to examine whether different types of signals provide complementary information for distinguishing AI-generated from human-written text, especially in difficult borderline cases and across heterogeneous text genres.

2. Related Work

Generative AI text detection has been investigated through a broad range of methods, from lightweight lexical and stylometric approaches to transformer-based architectures. Since our system combines lexical, linguistic, probabilistic, and semantic approaches, this section reviews prior work along these main methodological directions.

A first line of work focuses on statistical, lexical, and stylometric approaches [10, 11, 12]. These methods represent texts through explicit features such as lexical distributions, vocabulary richness, punctuation patterns, readability measures, and syntactic properties. For example, Lorenz et al. [10] showed that TF-IDF representations combined with logistic regression or SVM classifiers remain highly competitive despite their simplicity. Similarly, Jimeno-Gonzalez et al. [11] combined lexical and linguistic features with TF-IDF representations to build a lightweight and interpretable detector, while Seeliger et al. [12] proposed a signal-processing-inspired method based on word correlations. Overall, these

approaches are computationally efficient and relatively transparent, but their reliance on surface-level patterns may limit robustness to unseen generators, paraphrasing, or domain shifts.

A second line of work investigates zero-shot and probabilistic detectors [13, 14, 15]. These approaches exploit statistical properties of language model outputs, including perplexity, entropy, surprisal, and token-rank distributions, based on the assumption that AI-generated text tends to be more statistically regular than human-written text. One influential approach is Binoculars [14], which compares probability distributions produced by two language models to identify machine-authored text. Building on this idea, Tavan and Najafi [13] proposed BinocularsLLM, an ensemble combining probabilistic signals with neural classifiers, achieving first place in PAN 2024 [23]. Although these detectors require little or no task-specific training data, they can be sensitive to paraphrasing, prompt variation, and restricted access to model probabilities [15]. More recent research increasingly focuses on transformer-based architectures fine-tuned for text classification [16, 17, 18, 19]. These systems rely on contextual semantic representations learned by pretrained language models such as BERT, RoBERTa, DeBERTa, and Qwen. Representative approaches include chunk-based BERT classification [16], ModernBERT detectors with custom optimization strategies [17], DeBERTa-v3 with regularization and augmentation techniques [18], and large Qwen3-based detectors achieving state-of-the-art performance in PAN 2025 [19, 24]. Overall, transformer-based systems generally achieve stronger predictive performance because they capture contextual and discourse-level information, although they remain computationally expensive and difficult to interpret.

Another research direction explores representation learning and contrastive learning techniques [25, 26, 27]. These approaches aim to learn robust and transferable semantic representations through contrastive objectives, genre-aware embeddings, or multi-task learning architectures. Overall, they reflect a broader shift toward improving cross-domain robustness and representation generalization rather than relying exclusively on surface-level textual features.

Some studies have attempted to combine the strengths of different detection paradigms through hybrid and ensemble systems. These works provide evidence that heterogeneous signals can capture complementary properties of AI-generated text and improve robustness compared to single-model approaches. Kumar et al. [20] proposed a hybrid DistilBERT-based framework that combines semantic embeddings with handcrafted stylometric and entropy-based features. Marchitan et al. [21] explored multiple strategies for leveraging frozen and fine-tuned LLM embeddings together with machine learning classifiers such as logistic regression, SVM, XGBoost, and LightGBM. Their experiments showed that ensembles of classical classifiers operating on strong semantic embeddings can achieve performance comparable to fully fine-tuned transformers while requiring fewer computational resources. Similarly, Zaidi et al. [22] combined transformer architectures with data augmentation strategies, including back-translation and synonym replacement, to improve robustness in collaborative authorship detection. Our work builds on this direction by integrating a broader range of complementary lexical, linguistic, probabilistic, and semantic signals within a unified and partially interpretable ensemble framework.

Overall, the literature shows a clear evolution from lightweight stylometric methods toward increasingly sophisticated transformer-based and hybrid architectures. Statistical and zero-shot approaches offer practical advantages because of their efficiency and interpretability, while transformer-based systems often achieve stronger predictive performance. At the same time, recent work increasingly emphasizes robustness, domain adaptation, mixed-authorship detection, and explainability. Hybrid approaches that combine lexical, statistical, linguistic, and semantic signals therefore appear particularly promising, since they can offer a balanced trade-off between interpretability, computational efficiency, and predictive performance. However, fewer works explicitly investigate how different categories of signals complement each other within heterogeneous ensemble architectures.

In this work, our goal is to investigate whether heterogeneous and partially interpretable systems can support robust detection while providing better insight into subsystem behavior.

3. Proposed Framework

The proposed architecture is motivated by the assumption that robust AI-generated text detection requires evidence from multiple representational levels, especially under adversarial and style-mimicking conditions. As shown in Figure 1, our system combines four heterogeneous subsystems that capture complementary properties of machine-authored text, i.e. lexical regularities, linguistic structure, probabilistic smoothness, and semantic coherence. The **lexical subsystem** captures surface-level regularities

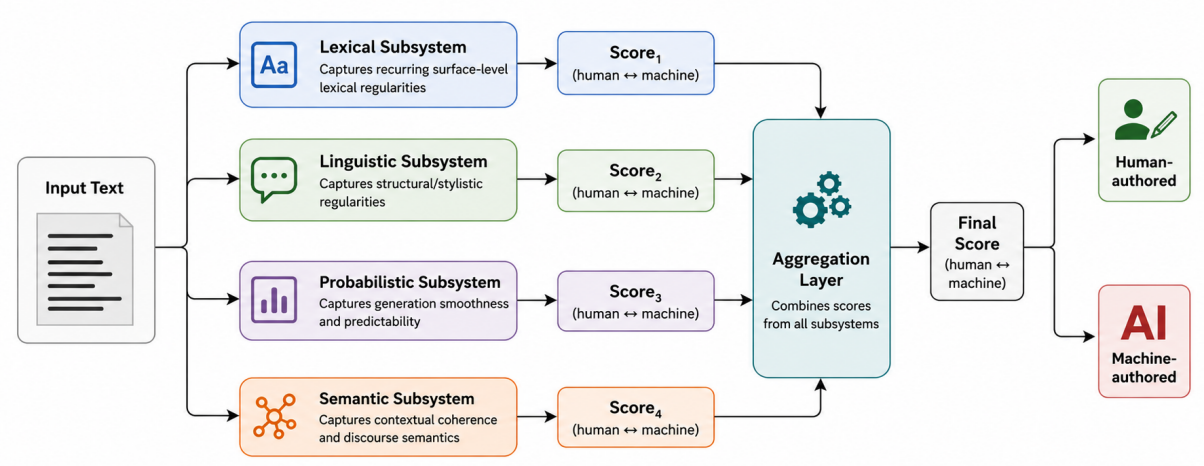

Figure 1: Overview of the proposed multi-signal ensemble for robust and interpretable AI-generated text detection.

in word use. For example, it can detect repeated expressions, frequent discourse markers such as "moreover" or "in conclusion", and unusually uniform lexical choices.

These signals help identify recurring wording patterns that may be associated with machine-authored text. The **linguistic subsystem** captures structural and stylistic regularities beyond individual words. For example, it can detect regular sentence lengths, repeated sentence openings, limited punctuation variation, and consistently formal writing patterns. These signals provide evidence about the organization and style of the text. The **probabilistic subsystem** captures generation smoothness and token predictability. For example, it can detect texts where many tokens are highly expected from the preceding context, or where uncertainty remains low across the text. These signals provide evidence about statistical regularities in the generation process. The **semantic subsystem** captures contextual coherence and discourse-level meaning. For example, it can detect very smooth topic transitions, repeated reformulations of the same idea, or strong semantic similarity between adjacent sentences. These signals provide evidence about the semantic flow and coherence of the text.

Together, these subsystem-level predictions provide complementary evidence for the final aggregation layer, while also allowing the classification decision to be inspected in terms of the type of signal that supported it.

3.1. Subsystem Instantiation

This section describes how we developed each subsystem and explains how each contributes complementary evidence to the final prediction.

Lexical subsystem. The lexical subsystem is implemented as a sparse text classifier based on TF-IDF representations extracted from word and character n-grams. We use word n-grams from 1 to 2 and character n-grams from 3 to 5, and classify the resulting feature vectors using logistic regression with L2 regularization. This design allows the subsystem to capture surface-level lexical and stylistic regularities associated with machine-authored text. Word n-grams capture recurring lexical patterns and phrases, while character n-grams capture punctuation usage, subword patterns, and stylistic cues that may not

be visible at the word level. logistic regression was selected based on our preliminary experiments using PAN 2026 training and validation sets [28, 29], where it consistently outperformed the other evaluated classifiers. This choice is also supported by previous work showing the effectiveness of logistic regression with sparse lexical representations for AI-generated text detection [10].

Linguistic subsystem. The second subsystem is based on handcrafted linguistic and stylometric features classified through logistic regression. To identify and design these features, we conducted a targeted literature review on linguistic patterns associated with AI-authored text. The final feature set was mainly inspired by the linguistic analysis proposed by Tercon and Dobrovoljc [30], which examines textual properties that distinguish AI-generated from human-written text. The subsystem extracts features from four main categories:

1. lexical repetition and diversity,
2. function word distributions,
3. syntactic and structural information,
4. domain and discourse style.

The first category models lexical diversity and repetition patterns. For example, the subsystem measures repeated unigram and bigram ratios, as well as lexical diversity metrics such as the number of unique tokens. These features are intended to capture highly regular lexical patterns, limited vocabulary variation, and repetitive phrasing, which may occur in AI-generated texts. During experimentation, they were useful for identifying texts with unusually stable word choices and recurring expressions.

The second category models stylistic behavior through function word distributions and stance-related expressions. For example, the subsystem measures the ratio of personal pronouns, such as “I”, “me”, “you”, “mine”, and “your”, together with the frequency of stance expressions, such as “perhaps”, “likely”, and “certainly”. It also captures epistemic markers, including expressions such as “I think”, “in my opinion”, and “I feel”. These features are intended to capture how explicitly a text expresses perspective, uncertainty, certainty, or interpersonal orientation. Such signals can be useful because human-written and AI-generated texts may differ in how they use personal reference, hedging, and stance-taking language.

The third category captures syntactic and structural information. Specifically, the subsystem measures sentence length variability, part-of-speech distributions, and punctuation frequency. These features are intended to capture how the text is structurally organized, including the regularity of sentence construction, the distribution of grammatical categories, and the use of punctuation marks. Such signals can help identify texts with unusually uniform sentence structures or highly regular syntactic patterns.

The final category models discourse organization and genre-related stylistic properties. These features were introduced after a qualitative analysis of frequent validation errors. Human-written news articles were sometimes incorrectly classified as AI-generated because of their regular phrasing, formal organization, acronyms, quotations, and repetitive reporting style. Conversely, some AI-generated fiction texts were harder to detect because they contained dialogue, descriptive narration, and greater stylistic variability. To capture these genre-dependent patterns, we introduced features related to dialogue frequency, attribution verbs, acronym usage, discourse markers, and sentence-opening diversity. These features model higher-level stylistic and discourse properties, such as the use of direct speech, reporting structures, repeated discourse transitions, and variation in sentence openings.

We intentionally selected handcrafted features because they remain interpretable. Each feature corresponds to a meaningful linguistic property that can be directly analyzed during qualitative error analysis and subsystem comparison. logistic regression was again selected because the feature space is relatively small and structured.

Probabilistic subsystem. The third subsystem is a probabilistic detector based on the Binoculars framework, combined with additional probabilistic features. For each input text, we extract the Binoculars score, mean and variance of surprisal, mean entropy, and mean log-rank. The resulting feature

vector is then used as input to a logistic regression classifier, which predicts whether the text is AI-generated. These probabilistic features provide information about token predictability and generation regularity from a language-modeling perspective. The Binoculars score was included because it provides a zero-shot detection signal without requiring task-specific language model training. The additional probabilistic features were introduced to enrich the subsystem with complementary information about token uncertainty and probability distribution behavior. This subsystem complements the lexical and linguistic components because it operates at the probabilistic level rather than the lexical or structural level.

Semantic subsystem. The semantic subsystem is implemented using DistilBERT, which we fine-tuned on the PAN 2026 training data [28] for binary AI-generated text detection. We include this subsystem to capture contextual and semantic information associated with AI-authored text. While the lexical and linguistic subsystems focus on surface patterns and writing style, the semantic subsystem models how words and sentences relate to each other in context. DistilBERT was selected because it provides strong contextual representations while remaining computationally efficient. This choice is consistent with the goal of the ensemble, where the semantic subsystem is intended to contribute complementary semantic evidence without dominating the more interpretable subsystems. The subsystem outputs a probability score indicating the likelihood that the input text is machine-authored.

Aggregation layer. The outputs of the subsystems are finally combined through an aggregation layer based on logistic regression. Each subsystem is implemented as a logistic regression classifier over its own feature representation and produces a probability score indicating whether the input text is AI-generated. The scores produced by the lexical, linguistic, probabilistic, and semantic subsystems are then used as input features for the final aggregation layer. This layer learns how to combine subsystem-level evidence and produces a single final probability score for each input text.

The motivation behind this ensemble design is that the usefulness of each subsystem may depend on the type of text being analyzed. Sparse lexical models provide strong global performance, while linguistic, probabilistic, and semantic subsystems may provide complementary evidence in difficult cases, such as highly formal institutional writing or AI-generated fiction with a human-like narrative structure. We selected logistic regression to preserve interpretability and facilitate the analysis of subsystem interactions. Since the aggregation layer operates on a small number of subsystem scores rather than high-dimensional raw features, a linear aggregation strategy is appropriate and allows direct inspection of subsystem contributions through model weights. This choice keeps the aggregation model simple and reduces the risk of overfitting compared to more complex aggregation strategies.

The final architecture therefore combines lexical, linguistic, probabilistic, and semantic evidence within a heterogeneous and partially interpretable ensemble framework for robust AI-generated text detection. In addition to the complete ensemble, we also designed several sub-ensembles combining specific subsets of the subsystems. The goal of these configurations was to analyze the individual contribution and complementarity of different feature families and to evaluate whether simpler combinations could achieve competitive performance with lower complexity.

4. Experimental Setup

Our empirical evaluation is conducted in the context of the PAN 2026 Voight-Kampff Generative AI Detection task [31], which focuses on distinguishing human-authored from machine-authored text under challenging generation conditions. The dataset provided by the organizers includes human-authored texts and AI-generated texts produced by multiple LLMs. The texts cover heterogeneous genres, including fiction, news, and essays. This genre diversity makes the task challenging because each genre exhibits different stylistic and structural properties.

4.1. PAN 2026 Dataset Analysis

Before designing the final architecture, we performed a preliminary analysis of the PAN 2026 training and validation data to better understand the characteristics of the detection task and identify which types of signals could be most informative.

We observed that different categories of texts exhibited substantially different stylistic properties. Fiction texts frequently contained dialogue, descriptive narration, and higher lexical variability, while news texts often exhibited highly regular institutional phrasing, acronyms, quotations, and formulaic reporting structures. Essays instead tended to present more stable argumentative organization and personal writing styles. The lexical regularities observed in many AI-generated texts motivated the inclusion of a sparse TF-IDF lexical subsystem. At the same time, the presence of genre-dependent stylistic variation motivated the introduction of handcrafted linguistic and domain-style features capable of modeling dialogue structure, discourse organization, sentence variability, and institutional writing patterns. The probabilistic subsystem was introduced to capture token predictability and generation smoothness, while the semantic transformer subsystem was designed to model higher-level contextual information that may not be fully captured by lexical or structural features alone.

Overall, the preliminary analysis suggested that a single detector family may not be sufficient to handle the heterogeneity of the dataset. For this reason, we adopted a multi-signal architecture that combines lexical, linguistic, probabilistic, and semantic evidence. This design aims to improve robustness, particularly in difficult borderline cases and across heterogeneous genres and generators.

5. System Analysis on the Validation Set

5.1. Subsystem Performance on the Validation Set

Table 1 reports the validation performance of the four standalone subsystems and the final aggregation layer, including F1, ROC AUC, accuracy, false positives, and false negatives. Overall, most subsystems achieve strong standalone performance, except for the probabilistic subsystem, which performs substantially lower than the others. The lexical subsystem achieves strong performance, indicating that lexical and orthographic regularities remain highly discriminative for AI-generated text detection. The semantic subsystem performs similarly, with fewer total errors, suggesting that contextual semantic modeling provides useful information beyond surface-level lexical patterns. The linguistic subsystem performs lower than both systems, but remains strong given its reliance on interpretable handcrafted features rather than large-scale lexical or semantic representations. The probabilistic subsystem is substantially weaker than the other detectors and produces a large number of false negatives. This indicates that probabilistic smoothness and token predictability alone are insufficient for robust standalone detection across heterogeneous genres and generators. Finally, the aggregation layer achieves the best overall performance. In particular, it reduces the number of false negatives compared to the lexical subsystem while maintaining a low number of false positives.

Table 1
Performance of the standalone subsystems and the aggregation layer on the validation set.

System	F1	ROC AUC	Accuracy	FP	FN
Lexical	.994	.999	.991	12	17
Linguistic	.978	.994	.971	45	58
Probabilistic	.848	.888	.814	207	459
Semantic	.994	1.000	.993	17	9
Aggregation layer	.995	1.000	.993	13	11

Table 2
Genre-level F1 scores for the subsystems and the aggregation layer on the validation set.

Genre	Lexical	Linguistic	Probabilistic	Semantic	Aggregation layer
Essays	.997	.984	.842	.995	.998
Fiction	.992	.990	.860	.995	.994
News	.993	.961	.837	.993	.994

5.2. Subsystem Behavior Across Genres

We further analyze how subsystem behavior varies across textual genres. In particular, we investigate whether specific detector families are more effective for particular genres and whether subsystem complementarity depends on the type of text being analyzed. To this end, we divide the validation set into three genres, namely essays, fiction, and news. For each genre, we evaluate the predictions produced by each subsystem and by the final ensemble independently.

Table 2 presents the genre-specific results, showing that subsystem effectiveness varies across text types. The lexical and semantic subsystems remain consistently strong across genres, while the linguistic subsystem shows greater variation, with stronger performance on fiction and lower performance on news. The probabilistic subsystem performs lowest across all genres, indicating that token predictability is not sufficient as a standalone signal in this setting. The aggregation layer achieves the best performance on essays and news, while the semantic subsystem performs best on fiction. Overall, these results suggest that aggregation can improve robustness, although its benefit depends on the genre.

5.3. Feature Analysis of the Lexical and Linguistic Subsystems

We analyze the lexical and linguistic subsystems on the validation set to better understand which features contribute most strongly to their predictions. This analysis uses the coefficients learned by the

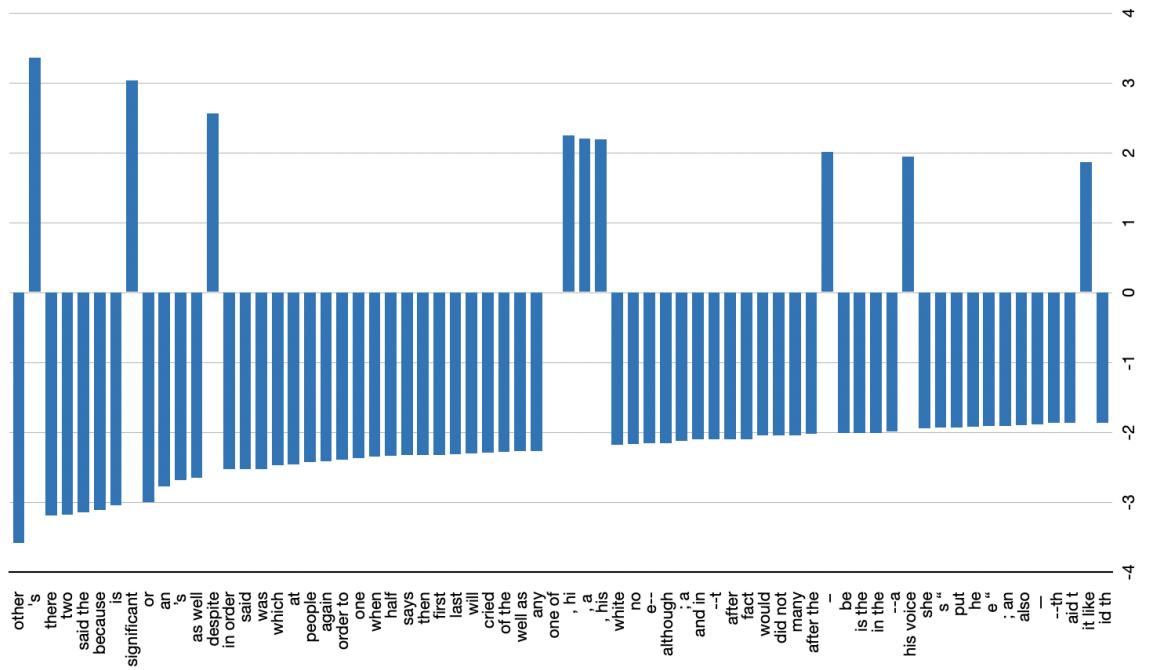

Figure 2: Logistic regression coefficients for word and character n-gram features in the lexical subsystem. Positive values indicate features associated with AI-generated text, while negative values indicate features associated with human-written text.

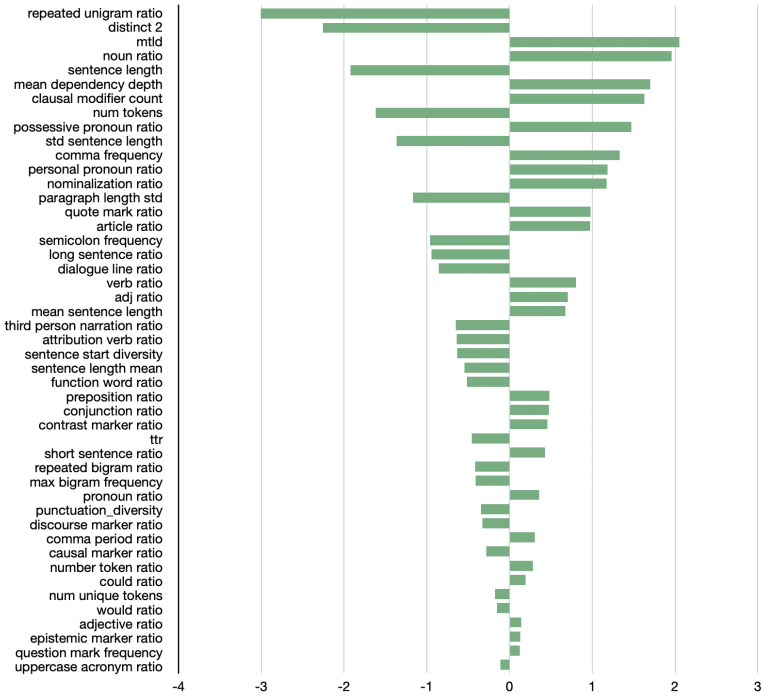

Figure 3: Logistic regression coefficients for handcrafted linguistic features. Positive values indicate features associated with AI-generated text, while negative values indicate features associated with human-written text.

corresponding logistic regression models. For the lexical subsystem, we inspect the most informative TF-IDF features extracted from word unigrams and bigrams, and character 3–5-grams. For the linguistic subsystem, we inspect the coefficients assigned to the handcrafted linguistic features. In both cases, positive coefficients indicate features that push the prediction toward AI-generated text, while negative coefficients indicate features that push the prediction toward human-written text.

Figure 2 shows the most influential lexical features. The results indicate that the lexical subsystem relies on both word-level expressions and character-level patterns. Several highly weighted features correspond to common lexical expressions, punctuation patterns, and character n-grams. This suggests that the sparse lexical subsystem captures not only word choice, but also orthographic and stylistic regularities. At the same time, some influential features appear to reflect genre-specific patterns, especially narrative or dialogue-related expressions, indicating that the lexical subsystem may partly encode genre-dependent surface cues.

Figure 3 shows the most influential handcrafted linguistic features. Positive coefficients are associated with features such as MTLD, noun ratio, dependency depth, clausal modifier count, comma frequency, and nominalization ratio. These features suggest that some AI-generated texts are characterized by greater lexical regularity, more nominal or structurally organized writing, and specific punctuation patterns. Negative coefficients are associated with features such as repeated unigram ratio, number of distinct tokens, sentence length, sentence length variability, dialogue line ratio, attribution verb ratio, and sentence start diversity. These features suggest that human-written texts in the validation set often contain greater structural variation, more dialogue-related patterns, and more diverse sentence organization.

6. Results

This section reports the results obtained by our submitted systems on the test set of the Voight-Kampff Generative AI Detection task at PAN 2026. For a complete comparison with the other participating systems, we refer the reader to the official task overview paper [32].

Among our submitted systems, an ensemble system combining only the lexical and linguistic subsystems achieved the best performance (reported mean score of .797) on the test set and ranked 10th out of the 34 reported systems in the official evaluation. This result is clearly higher than the standalone lexical, semantic, and probabilistic systems, as well as the ensemble configuration of all systems. This outcome suggests that the lexical and linguistic subsystems provide the most effective combination of signals under the characteristics of the documents included in the test set. The lexical subsystem offers a strong baseline based on surface-level regularities, while the linguistic subsystem adds complementary stylistic and structural information. This interpretation is consistent with the validation analysis, where lexical features showed strong standalone performance and linguistic features provided meaningful additional evidence despite lower standalone scores. The weaker performance of the other configurations suggests that larger ensembles do not necessarily improve effectiveness. When additional components are not well aligned with the strongest signals, they may dilute rather than reinforce the final prediction.

Overall, the results indicate that the lexical linguistic ensemble provides the best trade-off between effectiveness, efficiency, and stability among our submitted systems, partly due to its relative simplicity.

7. Conclusions & Future Work

In this work, we addressed AI-generated text detection through a heterogeneous and partially interpretable ensemble architecture that combines lexical, linguistic, probabilistic, and semantic subsystems.

The validation analysis shows that lightweight feature families capture useful signals for AI-generated text detection, including lexical regularities, orthographic patterns, stylistic variation, and discourse-level organization. These signals can support simple and scalable systems that remain easier to analyze than more complex black-box detectors. Although the absolute test performance leaves room for improvement, our best submitted system was fully interpretable and relied on a compact combination of lexical and linguistic evidence. This result supports the motivation of our work, since detection systems used in contexts involving authorship, responsibility, and credibility should be both effective and inspectable.

Future work could investigate more advanced methods for modeling subsystem complementarity, especially in difficult genre-dependent cases where different detectors exhibit different failure patterns. This direction is motivated by our error analysis, which shows that weaker subsystems can still provide useful evidence for some difficult cases while introducing noise in others. A promising direction is the development of uncertainty-aware aggregation mechanisms that selectively exploit complementary subsystem signals only when the classification decision is ambiguous. Future work could also explore multi-criteria decision making methods as an alternative to the logistic regression aggregation layer [33]. By treating subsystem scores as decision criteria, such methods could provide a fully interpretable score aggregation mechanism and allow us to examine whether transparent aggregation can improve performance while making the final classification easier to explain. Finally, future work should study how detector behavior changes across new language models, evolving prompting strategies, and different forms of text obfuscation.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This paper was supported by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly for paraphrasing, rewording, grammar checking, and spelling correction. Further, the authors used GPT-5.5 for figure 1 in order to: Generate a new image based on a draft image we created using draw.io, with the goal of improving content alignment and overall presentation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Liang, Y. Zhang, M. Codreanu, J. Wang, H. Cao, J. Zou, The widespread adoption of large language model-assisted writing across society, Patterns 6 (2025).

[2] M. A. Yeo, Academic integrity in the age of artificial intelligence (ai) authoring apps, Tesol Journal 14 (2023) e716.

[3] S. Peña Fernández, K. Meso Ayerdi, A. Larrondo Ureta, J. Díaz Noci, Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media, Profesional de la información 32 (2023).

[4] M. Chakraborty, S. T. I. Tonmoy, S. M. Zaman, S. Gautam, T. Kumar, K. Sharma, N. Barman, C. Gupta, V. Jain, A. Chadha, et al., Counter turing test (ct2): Ai-generated text detection is not as easy as you may think-introducing ai detectability index (adi), in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 2206–2239.

[5] J. Bevendorff, M. Wiegmann, E. Richter, M. Potthast, B. Stein, The two paradigms of llm detection: Authorship attribution vs authorship verification, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 3762–3787.

[6] D. La Barbera, G. Milanese, G. Peikos, G. Pasi, M. Viviani, et al., Beyond binary classification: ranking for information access in misinformation contexts, in: CEUR WORKSHOP PROCEEDINGS, volume 4121, CEUR-WS, 2025, pp. 1–7.

[7] J. D. Brüns, M. Meißner, Do you create your content yourself? using generative artificial intelligence for social media content creation diminishes perceived brand authenticity, Journal of Retailing and Consumer Services 79 (2024) 103790.

[8] The Guardian, The new york times drops freelance journalist who used ai to write book review, 2026. URL: https://www.theguardian.com/books/2026/mar/31/the-new-york-times-drops-freelance-journalist-who-used-ai-to-write-book-review, accessed: 2026-05-21.

[9] Gizmodo, Ai detectors are inaccurate and freelance writers are getting fired anyway, 2024. URL: https://gizmodo.com/ai-detectors-inaccurate-freelance-writers-fired-1851529820, accessed: 2026-05-21.

[10] L. Lorenz, F. Z. Aygüler, F. Schlatt, N. Mirzakhmedova, Baselineavengers at PAN 2024: Often-forgotten baselines for llm-generated text detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 2761–2768.

[11] M. Jimeno-Gonzalez, E. Martínez-Cámara, N. Fernandez, P. Díaz-García, L. A. U. López, Team SINAI-INTA at PAN 2025: Uncovering machine generated text with linguistic features, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3698–3706. URL: https://ceur-ws.org/Vol-4038/paper_293.pdf.

[12] M. Seeliger, P. Styll, M. Staudinger, A. Hanbury, Human or not? light-weight and interpretable detection of ai-generated text, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3929–3941. URL: https://ceur-ws.org/Vol-4038/paper_319.pdf.

[13] E. Tavan, M. Najafi, Marsan at PAN: binocularsllm, fusing binoculars' insight with the proficiency of large language models for machine-generated text detection, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 2901–2912. URL: https://ceur-ws.org/Vol-3740/paper-281.pdf.

[14] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, CoRR abs/2401.12070 (2024). URL: https://doi.org/10.48550/arXiv.2401.12070. doi:10.48550/ARXIV.2401.12070. arXiv:2401.12070.

[15] A. J. Gutiérrez Megías, L. A. Ureña-López, E. Martínez Cámara, The influence of the perplexity score in the detection of machine-generated texts, in: R. Mitkov, S. Ezzini, T. Ranasinghe, I. Ezeani, N. Khallaf, C. Acarturk, M. Bradbury, M. El-Haj, P. Rayson (Eds.), Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK, 2024, pp. 80–85. URL: https://aclanthology.org/2024.nlpaics-1.10/.

[16] B. Huang, C. Zhong, K. Yan, Y. Han, Author authentication of generative AI based on bert by regularization method, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 2619–2626. URL: https://ceur-ws.org/Vol-3740/paper-241.pdf.

[17] Z. Liang, K. Sun, H. Cao, J. Luo, Z. Han, Text author classification: A modernbert approach with gradient loss function, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3775–3780. URL: https://ceur-ws.org/Vol-4038/paper_301.pdf.

[18] B. Li, H. Qi, K. Yan, Team bohan li at PAN: deberta-v3 with r-drop regularization for human-ai collaborative text classification, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3763–3769. URL: https://ceur-ws.org/Vol-4038/paper_299.pdf.

[19] D. Macko, mdok of kinit: Robustly fine-tuned LLM for binary and multiclass ai-generated text detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3819–3826. URL: https://ceur-ws.org/Vol-4038/paper_307.pdf.

[20] R. Kumar, A. Trivedi, O. Varshney, Voight-kampff AI detection sensitivity, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3734–3739. URL: https://ceur-ws.org/Vol-4038/paper_296.pdf.

[21] T. Marchitan, C. Creanga, L. P. Dinu, Unibuc - NLP at "voight-kampff" generative AI detection PAN 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3827–3841. URL: https://ceur-ws.org/Vol-4038/paper_308.pdf.

[22] S. A. Zaidi, H. T. Ahmed, S. A. Akbar, Z. Shakeel, F. Alvi, A. Samad, Team nexus interrogators at PAN: voight-kampff generative AI detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 4063–4072. URL: https://ceur-ws.org/Vol-4038/paper_335.pdf.

[23] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korencic, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein,

M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification condensed lab overview, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, Lecture Notes in Computer Science, Springer, 2024, pp. 231–259. URL: https://doi.org/10.1007/978-3-031-71908-0_11. doi:10.1007/978-3-031-71908-0_11.

[24] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-kampff generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Lecture Notes in Computer Science, Springer, 2025, pp. 388–411. URL: https://doi.org/10.1007/978-3-032-04354-2_21. doi:10.1007/978-3-032-04354-2_21.

[25] J. Liu, L. Kong, Z. Peng, F. Chen, Generative AI authorship verification based on contrastive-enhanced dual-model decision system, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3793–3802. URL: https://ceur-ws.org/Vol-4038/paper_304.pdf.

[26] J. Yang, K. Yan, Genre-aware contrastive learning for AI text detection: A roberta-based approach, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 4053–4058. URL: https://ceur-ws.org/Vol-4038/paper_333.pdf.

[27] A. Voznyuk, G. Gritsai, A. V. Grabovoy, Team advacheck at PAN: multitasking does all the magic, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 4007–4014. URL: https://ceur-ws.org/Vol-4038/paper_328.pdf.

[28] J. Bevendorff, M. Wiegmann, M. Potthast, B. Stein, Pan'25/26 generative ai detection: Voight-kampff ai detection sensitivity, 2025. URL: https://doi.org/10.5281/zenodo.14962653. doi:10.5281/zenodo.14962653.

[29] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, Overview of pan 2026: Voight-kampff generative ai detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[30] L. Tercon, K. Dobrovoljc, Linguistic characteristics of ai-generated text: A survey, CoRR abs/2510.05136 (2025). URL: https://doi.org/10.48550/arXiv.2510.05136. doi:10.48550/ARXIV.2510.05136. arXiv:2510.05136.

[31] J. Bevendorff, et al., Pan 2026: Voight-kampff generative ai detection, https://pan.webis.de/clef26/pan26-web/generated-content-analysis.html, 2026. Accessed: 2026-05-22.

[32] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[33] G. Peikos, G. Pasi, et al., A decision-theoretic framework to multidimensional relevance estimation, International Journal of Information Technology & Decision Making (2025) 1–43.